CSC8631_C1011323_EDA-Report

Sahej Sareen

02/12/2021

Introduction

The demand for cyber security professionals has increased significantly in the last decade. This is because there has been substantial increase in the use of technology such as advance cell phone technology, high speed internet accessibility etc., which has led to increase in the data that is being generated. Thereby, the need to protect this data from vulnerabilities, risks, failures and attacks from hackers is imperative.

The demand to build a cyber resilient strategy becomes a necessity to avoid reputation damages and protection from privacy invasion. There are now numerous programmes which prepare students to cope with the complex future of cyber security. Most of the courses related to cyber security are online and can be recorded or downloaded for further reference. These online courses access a network of world class academics and improve the courses handling. The Future Learn MOOC data set contains archetype-survey-responses, enrollments, leaving survey response, question response, statistical viewing of viewing, question response. The data recorded for the above observations are done in specific intervals.

In this project work the analysis is done on the statistical viewing of viewing data set. This data set contains varies parameters such as Title, total views, duration of the videos, location of videos where they were watched. Using the above data along with the guidance of Cross-Industry Standard Process for Data Mining (CRISP-DM) Methodology, exploratory data analysis is performed. The results obtained graphically and with the calculations helps gives an insight on the data which can then be used to solve the business problems.

Reproducibility

1. The meaning of reproducibility is the about running the same model n number of times but in return get the same output after every run which can be later modified to run investigation.
2. Project Template helps in the automating the process for us. Starting from preprocessing the data, organising the files in the folder, loading the R packages , loading all the data sets in the computers memory.
3. The benefits of this is that it results in faster loading of the files as it contains all the data in the cache folder, avoids problems between the code and the workspace code.
4. During the processing data the complete data set is in taken into accountability as one whole unit as a single data frame. If there is any more files to added needs to be done manully.
5. Project writing becomes easier as the reported is generated automatically in R markdown without any hassle.

Analysis

The above data analysis is done based on the CRISP-DM methodology. The following steps in this methodology are as follows

a)Business Understanding

This is essential part of the project before proceeding ahead with the next step. This is because it is essential to understand what the business objective will be otherwise the half way through the project the goal will not be achieved and economical resources as well as time will be wasted.

b)Data Understanding

• In this step, there are 4 steps namely collect,Describe,explore, data quality. • The data collection is where we acquire the data. Describe is where we check the data to see if there any errors in it, also check its structure. Also to check if the data is satisfying our objective or not. • In data Exploring the data and submitting the initial findings of the data.
• Data quality to check if there are any blank spaces in or not.

c)Data preparation

Now after acquiring the data and at this step it is the final step before doing modelling on it. In this we we clean the data and maybe construct new columns or rows in it to help in the modelling section. Also in this we integrate two or more datas.

d)Modelling

In this step after data preparation we apply various models on it and test various outcomes on it and later assess the models. e)Evaluation In this step we evaluate the results obtained. This is the summary of all the steps previously worked on

f)development

In this process decision is taken to be worked on or not. If agreed upon then it is executed otherwise it will go for another iteration.

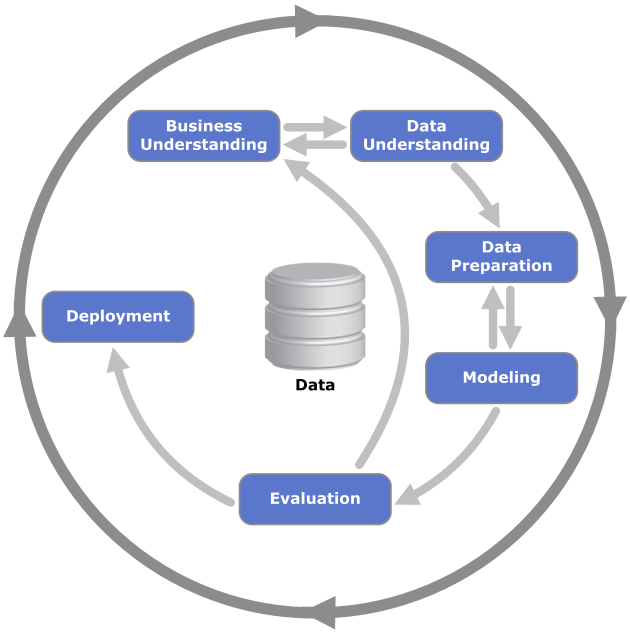

Figure 1: The CRISP-DM Model

Q1) Cycle 1- How are the total views effected by the Step position.

Reason for asking question-

The reason for asking this question is that some people might only want to study only specific and maybe they study that specific topic and leave thereby losing the viewership.If the institution knew about this

problem they might be able to segregate into smaller videos and make it more interesting to attract other students.

In this question we need to understand whether there is any relation between the views and the step position. This is the business understanding step in the CRISP-DM methodology. The first step before beginning with our analysis, the data needs to be cleaned and the columns must be structured.

Now comparing the columns of total views effected by the Step position in video_stats_3 which is for the month of September 2017. In order to find the relation between the total views and step position, the correlation function is applied. The value calculated is -0.7168596. This states that they are inversely proportionate to each other. Later the graph is plotted via ggplot() function. As per the graph we can conclude that the viewers decreases as they proceed through the course.

Comparing the columns in the video_stats4 as well as for video_stats5 by repeating the same procedure. The results observed by calculation for videos statistics 4 is -0.673233 and for video stats5 is -0.6636169 . Using ggplot()for plotting the graph for video_stats 4 and video_stats 5, it can be inferred that the viewership starts to decrease as the course progresses towards the end.

```
cortop1 <- cor(cyber.security.3_video.stats[,1],cyber.security.3_video.stats[,4])
cortop1
```

```
##               total_views
## step_position  -0.7168596
```

```
cortop2 <- cor(cyber.security.4_video.stats[,1],cyber.security.4_video.stats[,4])
cortop2
```

```
##               total_views
## step_position   -0.673233
```

```
cortop3 <- cor(cyber.security.5_video.stats[,1],cyber.security.5_video.stats[,4])
cortop3
```

```
##               total_views
## step_position  -0.6636169
```

Q2) Cycle 1 How many people who watched the videos have completely watched the videos till the end?

Reason for asking question-

To understand the number of peoples engagemnet of the students?

In this question we will be merging the data from the video_stats4 and video_stats5 i.e from November 2017 to February 2018. After combining the data files the correlation function is applied between columns total view and viewed onehundred percent to find the relation between them. From the calculated value the result obtained is 0.03268353. For plotting the graph the ggplot() function is used.

From the correlation of the total views and onehundred percent it can be stated that:

a) Whenever the total views increases not only the total learners watching 100 percent of video increases but also the percentage learners increases and vice versa.

b) This shows that if any video has good content people viewing it increases at a healthy rate which is stated through positive numerical correlation observed here.

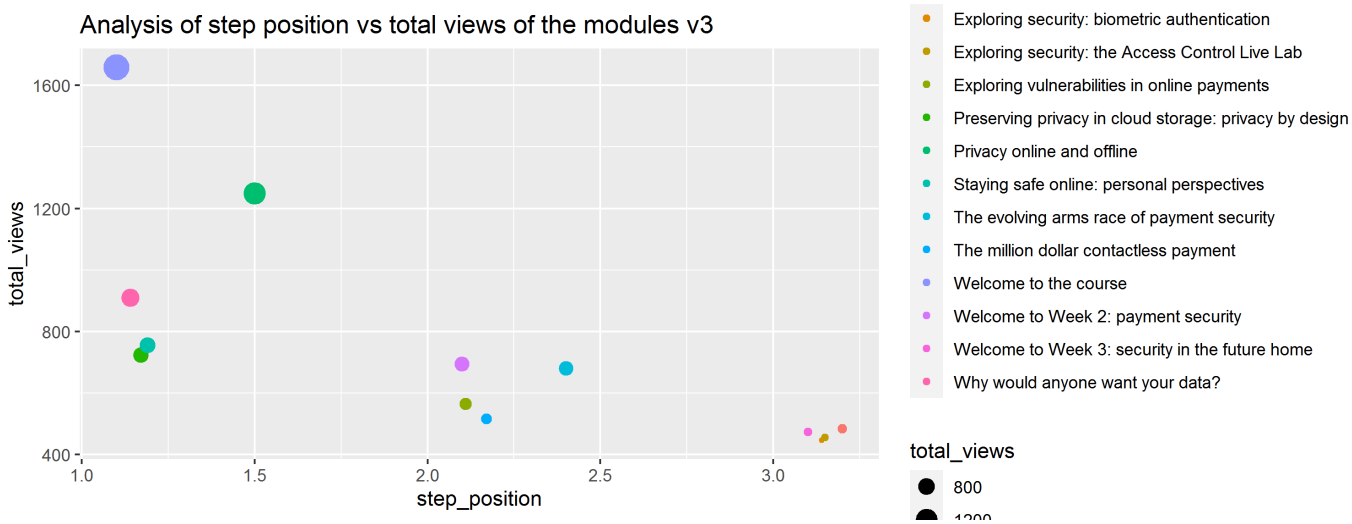

Figure 2: Analysis of step position vs total views of the modules v3

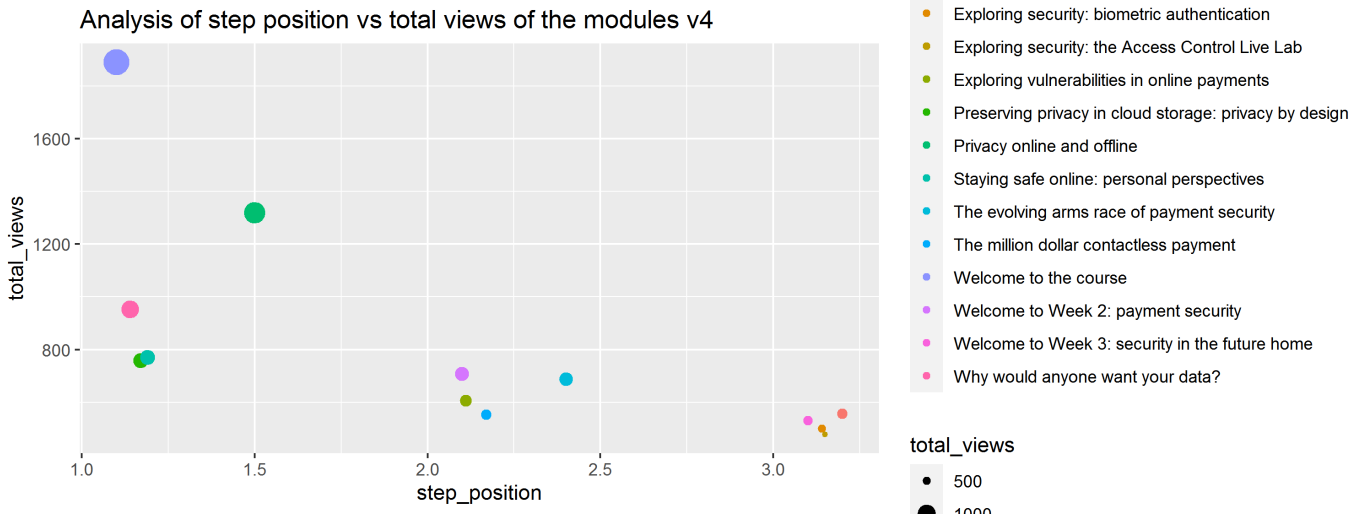

Figure 3: Analysis of step position vs total views of the modules v4

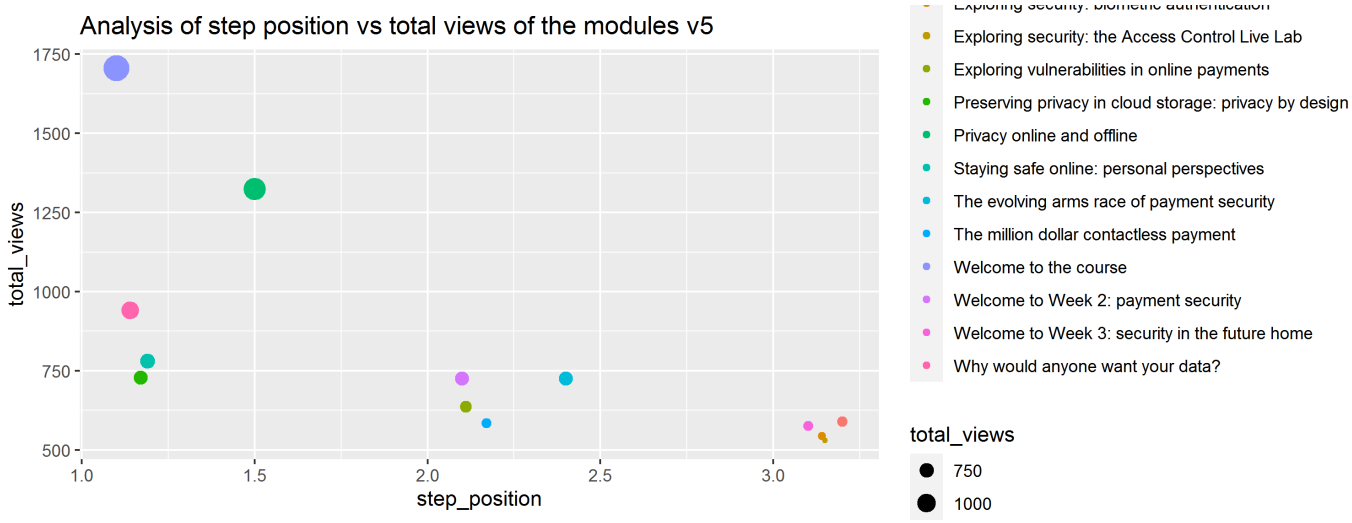

Figure 4: Analysis of step position vs total views of the modules v5

```
videostats4and5 <- rbind(cyber.security.4_video.stats, cyber.security.5_video.stats)
videostats4and5
```

```
## # A tibble: 26 x 28
##    step_position title                video_duration total_views total_downloads
##            <dbl> <chr>                         <int>       <int>           <int>
##  1           1.1  Welcome to the cour~            99        1890             192
##  2           1.14 Why would anyone wa~           362         952              99
##  3           1.17 Preserving privacy ~           241         757              86
##  4           1.19 Staying safe online~           348         770              84
##  5           1.5  Privacy online and ~           281        1318             144
##  6           2.1  Welcome to Week 2: ~            37         708              55
##  7           2.11 Exploring vulnerabi~           312         607              63
##  8           2.17 The million dollar ~            92         552              52
##  9           2.4  The evolving arms r~           426         688              68
## 10           3.1  Welcome to Week 3: ~            59         530              46
## # ... with 16 more rows, and 23 more variables: total_caption_views <int>,
## #   total_transcript_views <int>, viewed_hd <int>, viewed_five_percent <dbl>,
## #   viewed_ten_percent <dbl>, viewed_twentyfive_percent <dbl>,
## #   viewed_fifty_percent <dbl>, viewed_seventyfive_percent <dbl>,
## #   viewed_ninetyfive_percent <dbl>, viewed_onehundred_percent <dbl>,
## #   console_device_percentage <dbl>, desktop_device_percentage <dbl>,
## #   mobile_device_percentage <dbl>, tv_device_percentage <dbl>, ...
```

```
cortop4 <- cor(videostats4and5[,4],videostats4and5[,15])
cortop4
```

```
##             viewed_onehundred_percent
## total_views                0.03268353
```

Reason for asking question-

After understanding the engagement of the students, it is time to understand why the students have left the videos and what can be doen to avoid this problem. This will not only help in attracting newer students

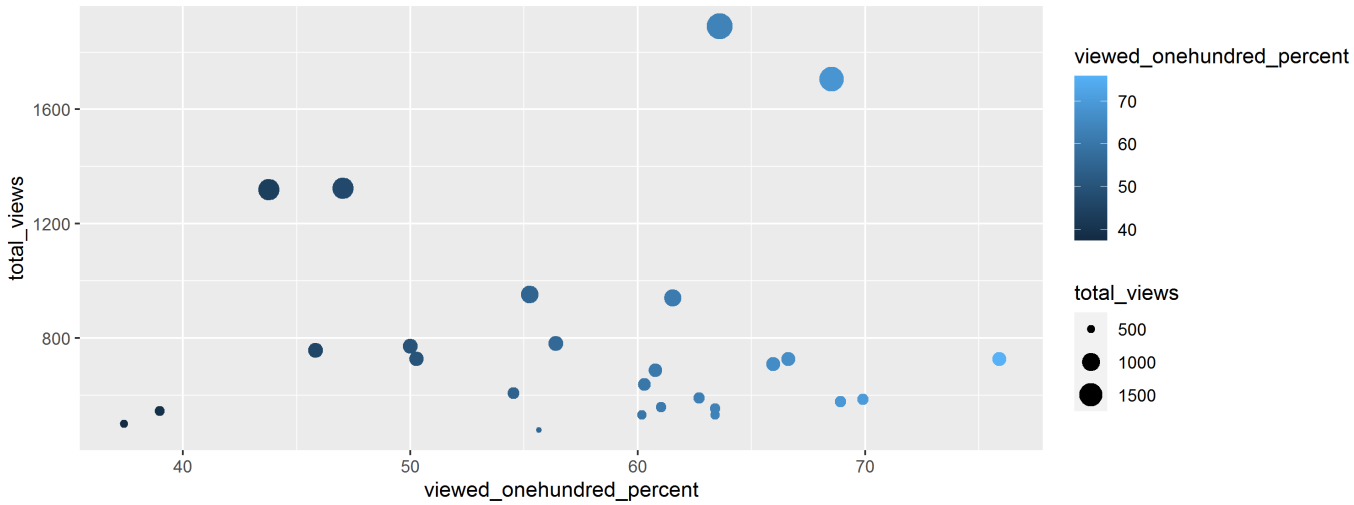

Figure 5: relation between total views and viewed onehundred percent v4 and v5

and might help reduce loads on the existing resources and can be tailored according to the comfortablity of the students.

Cycle 2 What is the least and most common reason to leave the course?

The data for this file is taken from the Leaving survey response. The data is transformed in data frame and the percentage of leaving reason variable is calculated. The data file for leaving response 4and5 are merged together. The data is then cleaned by omitting the blank columns.

There are some text in the heading of the columns which needs to be changed to ensure that the readers can understand what is written. This ensure that there is no error while performing our analysis. The output is divided into seven categories. They are as follows

a) I don't have enough time
b) Other
c) The course required more time than I realized
d) The course was too easy
e) The course was too difficult
f) The course wasn't what I expected
g) The course won't help me reach my goal

The following data is transfered from

1.From the follwoing data "The course wasnâ€™t what I expected"] to "The course wasn't what I expected"

2. From the follwoing data "I donâ€™t have enough time"] to "I don't have enough time"
3. From the follwoing data "The course wonâ€™t help me reach my goals"] to "The course won't help me reach my goals"

It can be understood from the graph that the pie chart shows that the majority of students chose "Other" as their reason for leaving the course (i.e. 30.09 percent). Thereby no conclusions can be drawn from this. Thereby selecting the next highest column that is they ' I don't have enough time' i.e 17.70%. The least percentage of the pie is the 'the course was too easy' i.e 7.96%.

Q3) Does the duration of the video effect the total viewership?

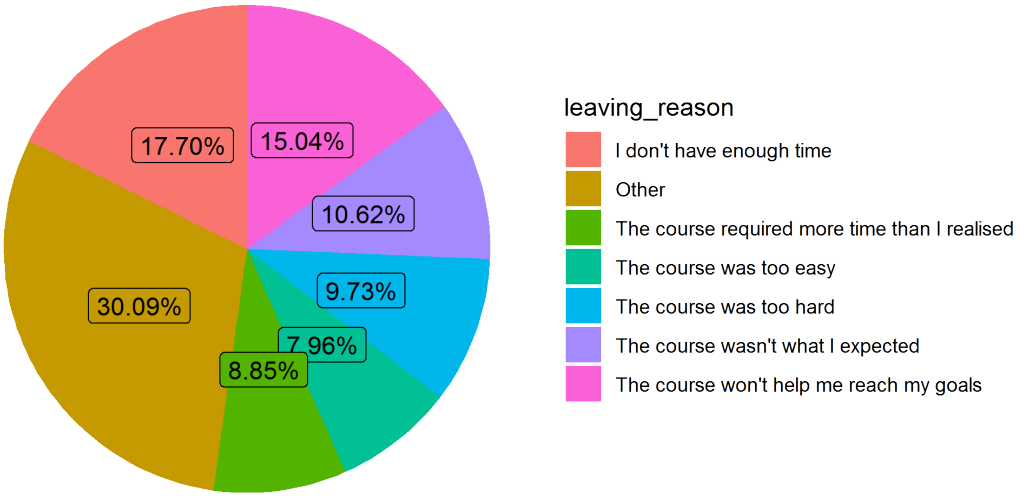

Figure 6: leaving reason

Reason for asking question-

This will in a way help in understanding the interest of the students and the videos can be shortened or students can be given breaks in between. Maybe for certain lectures can be lengthier to cover much more course and reduce the over course duration of the subject.

In this question we will be considering the column of the total views and duration of the video. In this case we will be considering the data for September 2017 November 2017 and February 2018. This will help us in performing the analysis and give us some insights of the relation. We will be using the correlation function to find out the relation between them. After performing the correlation function the calculated result for video_stats3 is -0.0532359.

The same procedure is repeated on the video_stats4 is -0.09594468. and the result for video_stats5 is -0.07226244. It can be said that the relation is inversely proportional to each other. After performing the correlation function a graph is plotted using the ggplot() function for the above mentioned data. As per the plot it can be seen that as video duration increase the total viewers dip in number. Figure7,8,9

```
cortop6 <- cor(cyber.security.3_video.stats[,1],cyber.security.3_video.stats[,4])
cortop6
```

```
##               total_views
## step_position  -0.7168596
```

```
cortop7 <- cor(cyber.security.4_video.stats[,1],cyber.security.4_video.stats[,4])
cortop7
```

```
##               total_views
## step_position   -0.673233
```

```
cortop8 <- cor(cyber.security.5_video.stats[,1],cyber.security.5_video.stats[,4])
cortop8
```

```
##                total_views
## step_position  -0.6636169
```

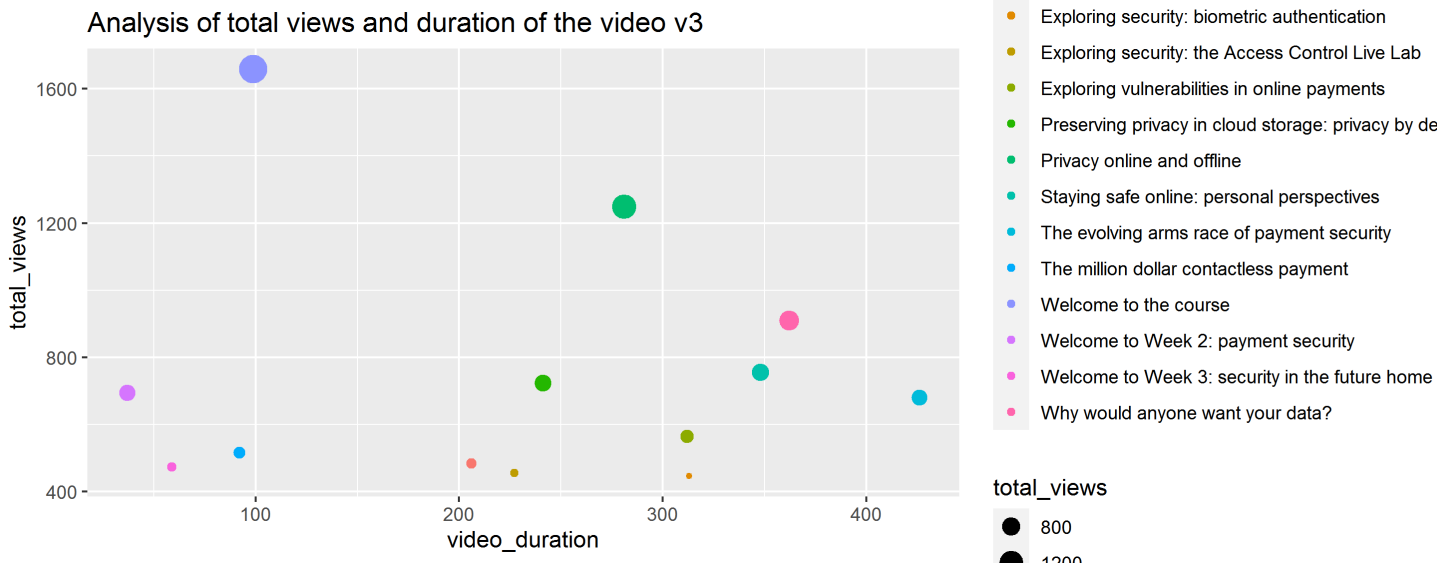

Figure 7: total views and duration of the video v3

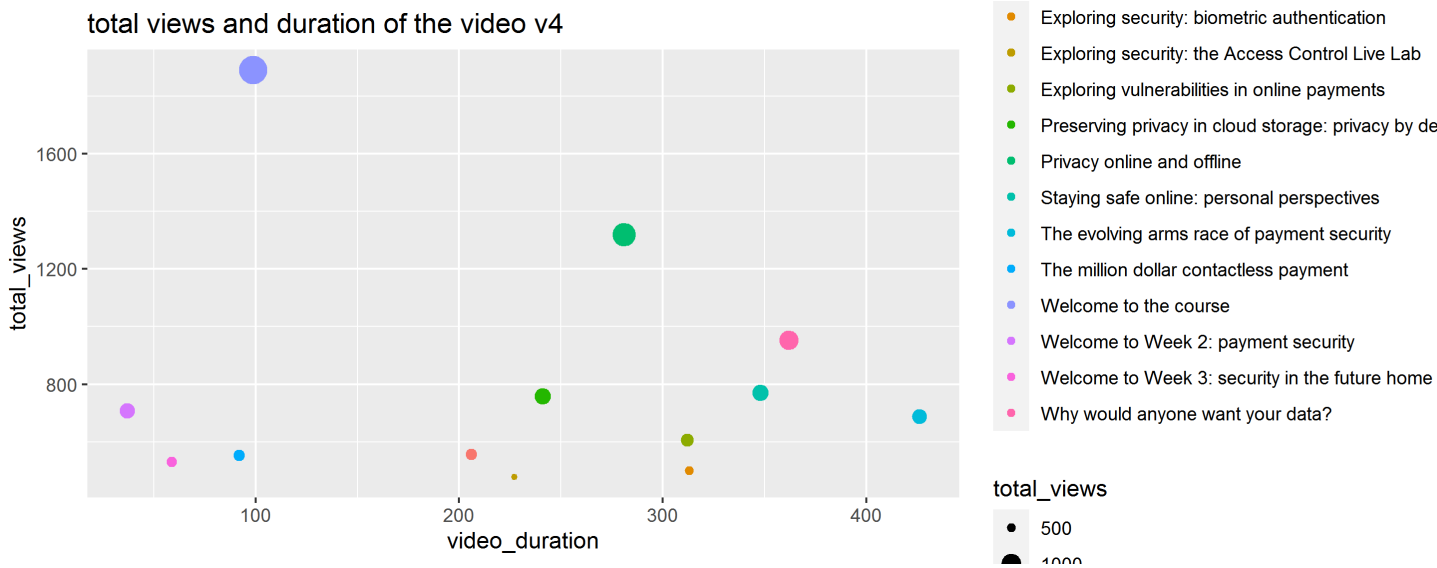

Figure 8: total views and duration of the video v4

Q4) Exploring other data such as step position and total viewed HD to find out any relation between them?

Reason for asking question-

This is because it can help the faculty to understand whether expensive cameras are recovered or can be shot on low end cameras. This might help in reducing the storage place for the videos and whther it is required to buy expensive shooting gear.

In this question we will be considering the step position and total viewed HD for our analysis. In this case we will be considering the data for the months of September 2017 November 2017 and February 2018. In

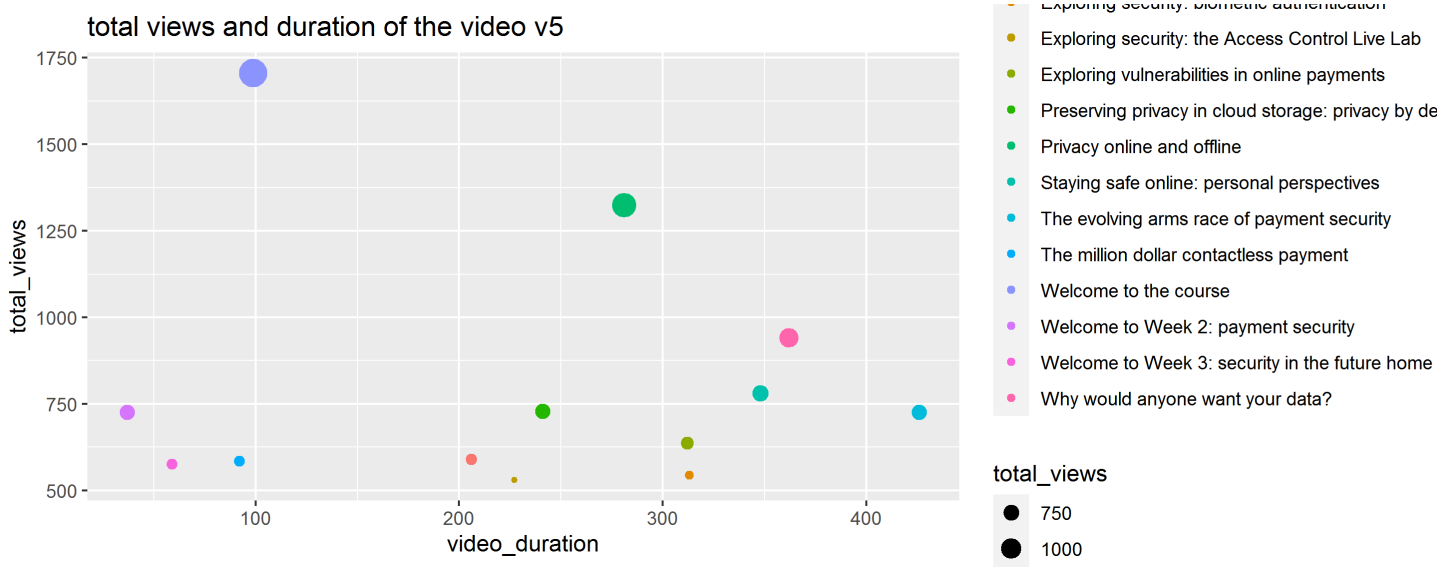

Figure 9: total views and duration of the video v5

this to find out the relation between the step position and total viewed hd we will be applying the cor() function.

The answers obtained at the end of the calculation of correlation for video_stats3 is -0.0532359. The same process is applied to the video_stats4 and video_stats5. The obtained for them are -0.06281232 and -0.04673555. This shows that the result is much closer to 0, this indicated that there is no correlation between them.

Using ggplot() for plotting the graph for all the three data. We can infer that there is the HD quality videos have no effect on the student rather it indicates the makers that the data must have substantial content and not depend on the HD quality of video to interest the students. Figure10,11,12

```
cortop9 <- cor(cyber.security.3_video.stats[,1],cyber.security.3_video.stats[,8])
cortop9
```

```
##                 viewed_hd
## step_position -0.08518971
```

```
cortop10 <- cor(cyber.security.4_video.stats[,1],cyber.security.4_video.stats[,8])
cortop10
```

```
##                 viewed_hd
## step_position -0.06281232
```

```
cortop11 <- cor(cyber.security.5_video.stats[,1],cyber.security.5_video.stats[,8])
cortop11
```

```
##                 viewed_hd
## step_position -0.04673555
```

Conclusion

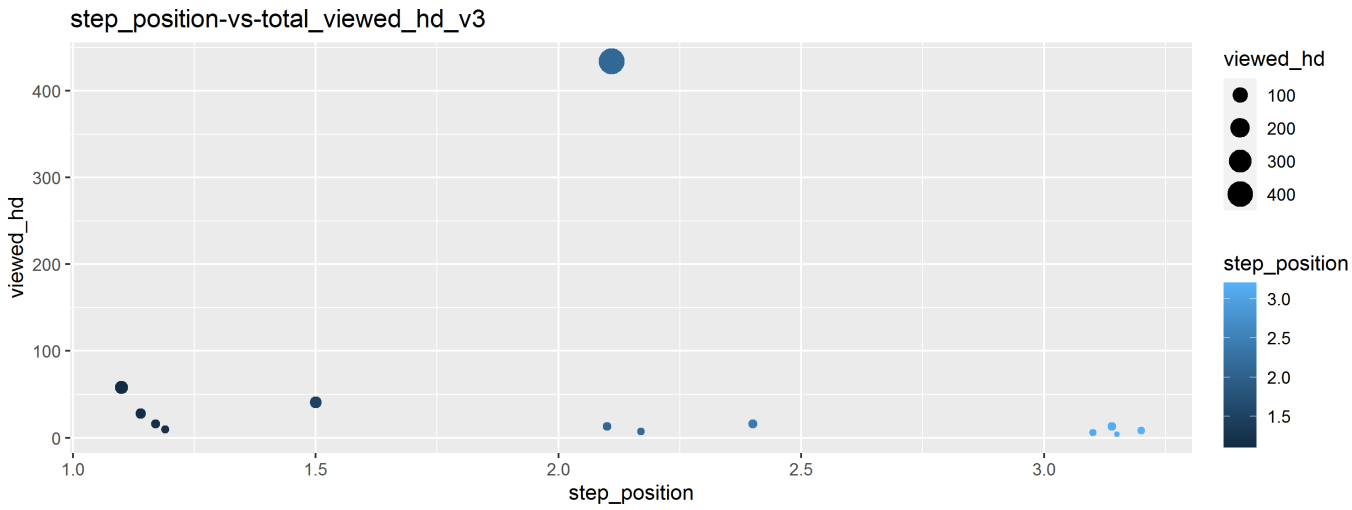

Figure 10: v3

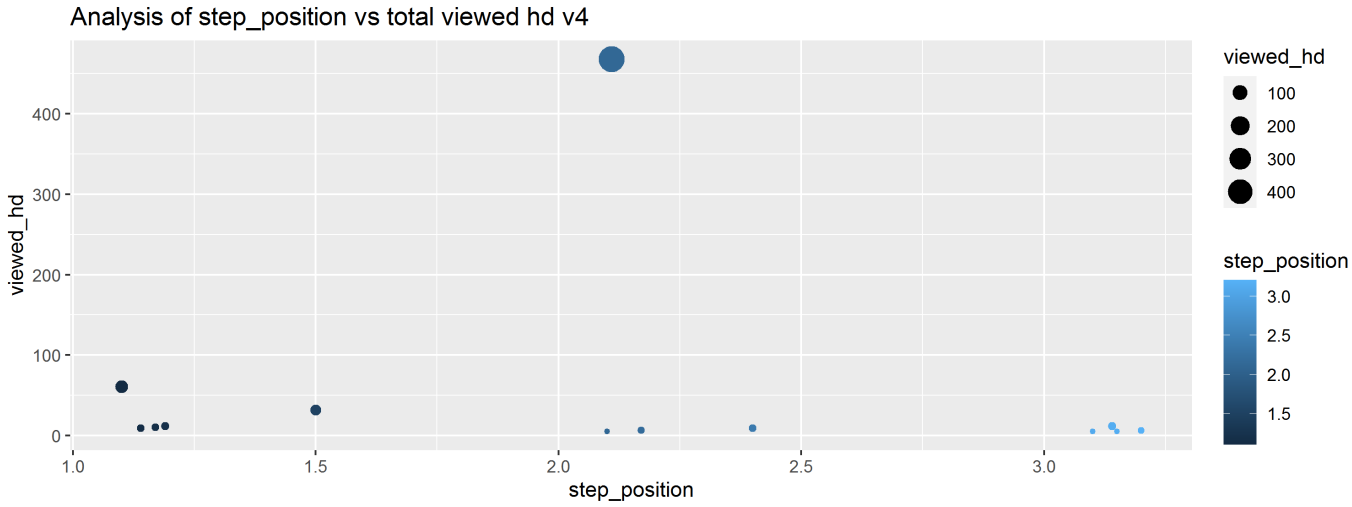

Figure 11: v4

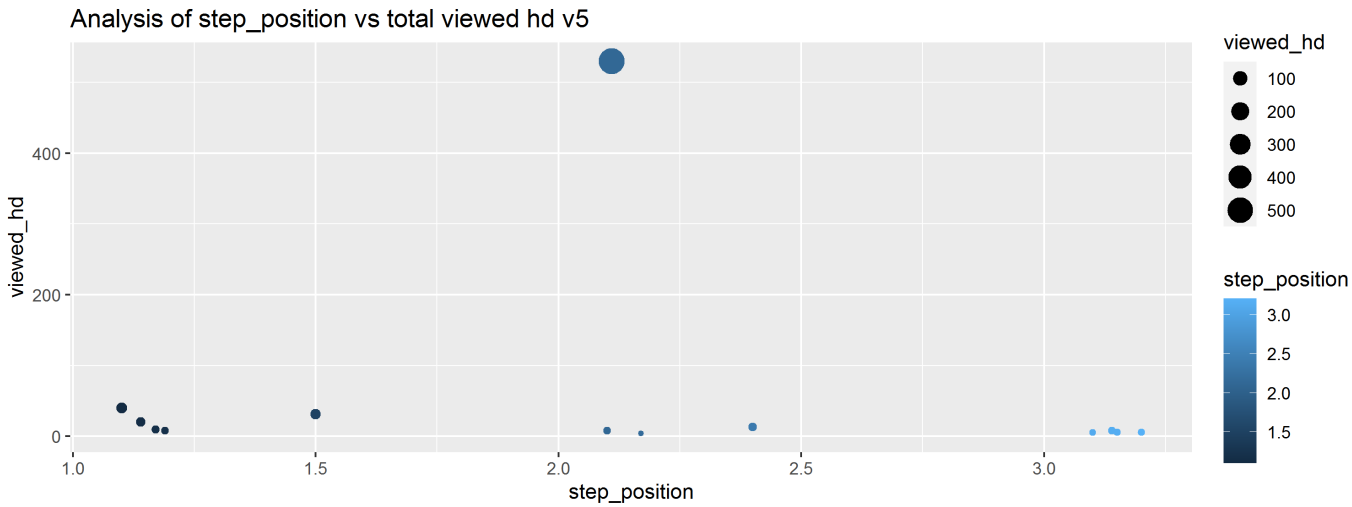

Figure 12: v5

1.After exploring moderate amount of the data it can inferred that sub titles of the subject does not play much role in the viewers leaving the course early on. The viewers interested sat through the complete course whereas the reason for the people leaving was expressed that they did not have enough time to study the course.

2.The other concern that the duration of videos made the viewers turn away from the course. This has some sort of connection that since they do not have much and the which makes them not watch the videos even if they download.

3.The last data that was explored between the step position and HD video has little to no connection from the graph. Thereby, it can be inferred that the content of the video matters and the clarity of the video.